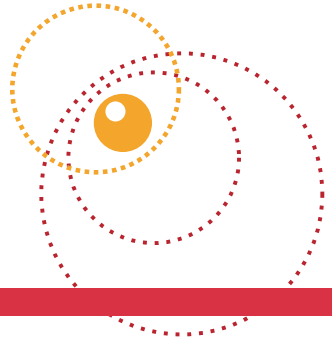


Context tree models for source coding

Toward Non-parametric Information Theory

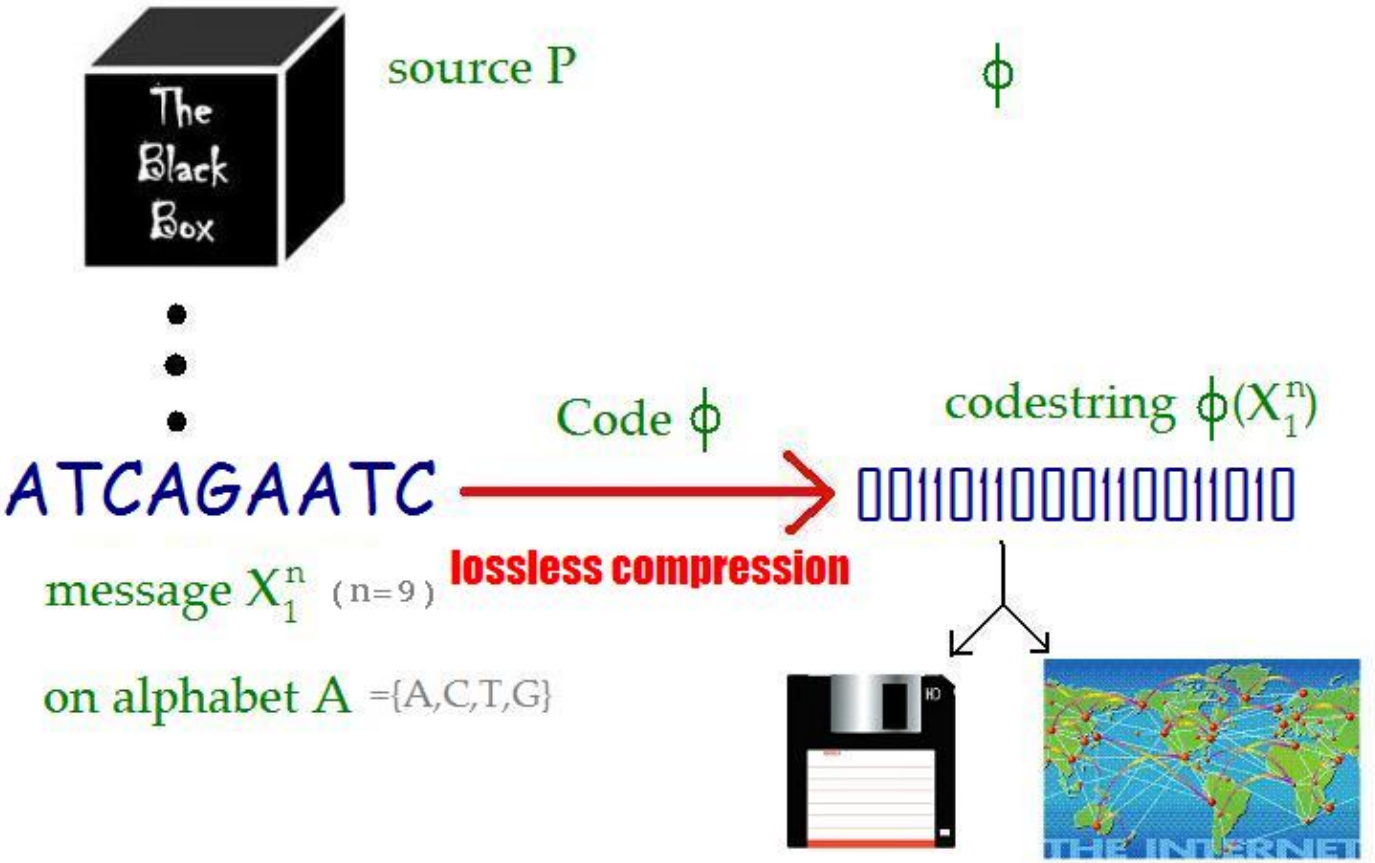
August, 06 2009

Aurélien Garivier

Outline

Aurélien Garivier
<garivier@enst.fr>

Data Compression : Shannon Modelization



Data Compression : Shannon Modelization

Source P

= stationnary process on **Alphabet A** ={A,C,T,G}

Code $\phi_n : A^n \longrightarrow \{0,1\}^*$

ATCAGAATC → 001101100011001101O

Message $X_1^n (n=9)$

lossless compression

Winzip, compress, etc.

Codestring $\phi_n(X_1^n)$

Goal : minimize average codelength

$$E_P\left[|\phi(X_1^n)|\right]$$

Outline

Aurélien Garivier
<garivier@enst.fr>

The Problem

- Finite *alphabet* A ; ex : $A = \{a, c, t, g\}$
- Set of all finite binary sequences : $\{0,1\}^* = \bigcup_{n \geq 0} \{0,1\}^n$
- If $x \in A^n$, its *length* is $|x| = n$
- log = base 2 logarithm
- Let X be a stochastic process on the finite alphabet A, with stationary ergodic distribution $\mathbb{P}$
- The marginal distribution of X_1^n is denoted by P^n
- For $n \in \mathbb{N}^*$, the *coding function* $\phi_n : A^n \to \{0,1\}^*$
- Expected code length :

$$\mathbb{E}_{\mathbb{P}} \left[|\phi_n(X_1^n)| \right]$$

Example : DNA Codes $A = \{a, c, t, g\}$

- Uniform concatenation code :

$$\phi_1 = \left| \begin{array}{cccc} a & c & t & g \\ 00 & 01 & 10 & 11 \end{array} \right|, \phi_n(X_1^n) = \phi_1(X_1) \dots \phi_1(X_n)$$

$$\Longrightarrow \forall x_1^n \in A^n, |\phi_n(X_1^n)| = 2n$$

- Improvement ? $\psi_1 = \left| \begin{array}{cccc} a & c & t & g \\ 0 & 1 & 01 & 10 \end{array} \right|$

  Problem : what is $\psi_1^{-1}(101)$: *ct* or *gc* ?

- *Uniquely decodable* codes :

$$\phi_n(X_1^n) = \phi_m(Y_1^m) \Longrightarrow m = n \text{ and } \forall i \in \{1, \dots, n\}, X_i = Y_i$$

- *Instant* codes = *prefix* codes :

$$\forall (a, b) \in A^2, \forall j \in \mathbb{N}, \phi_1(a)_{1:j} \neq \phi_1(b)$$

Kraft's Inequality

Theorem (Kraft '49, McMillan '56)
If ϕ_n is a *uniquely decodable* code, then

$$\sum_{x \in A^n} 2^{-|\phi_n(x)|} \leq 1$$

Proof for prefix codes : Let $D = \max_{x \in A^n} |\phi_n(x)|$ and for $x \in A^n$, let $Z(x) = \{w \in \{0,1\}^D : w_1^{|\phi(x)|} = \phi_n(x)\}$.

- $\#Z(x) = 2^{D-|\phi_n(x)|}$
- ϕ_n prefix $\implies \forall x, y \in A^n, x \neq y \implies Z(x) \cap Z(y) = \emptyset$
- Hence

$$\sum_{x \in A^n} N(x) = \sum_{x \in A^n} 2^{D-|\phi_n(x)|} \leq \#\{0,1\}^D = 2^D$$

Shannon's First Theorem

$\Longrightarrow$ if $\ell(x) = |\phi_n(x)|$, we look for

$$\min_{\ell} \sum_{x \in A^n} \ell(x) P^n(x) \text{ under constraint } \sum_{x \in A^n} 2^{-\ell(x)} \leq 1$$

Theorem (Shannon '48)

$$\mathbb{E}_{\mathbb{P}} \left[|\phi_n(x)| \right] \geq H_n(X) = \mathbb{E}_{\mathbb{P}} \left[-\log P^n(X_1^n) \right]$$

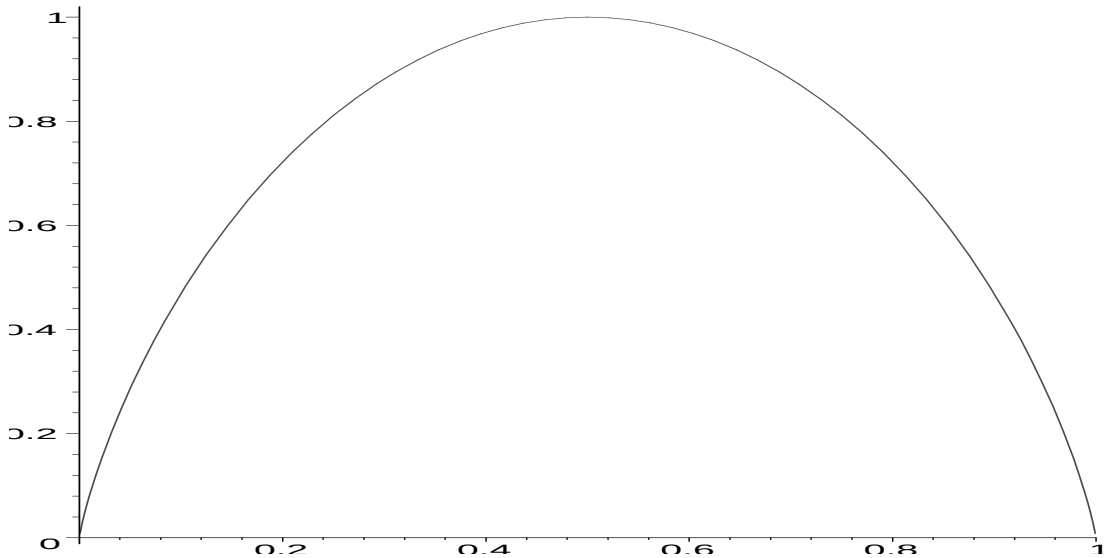
- Misleading notation : $H_n(X)$ is actually a function of P^n only, the marginal distribution of X_1^n
- Tight bound : there exist a code ϕ_n such that

$$\mathbb{E}_{\mathbb{P}} \left[|\phi_n(x)| \right] \leq H_n(X) + 1$$

Example : binary entropy

$$A = \{0, 1\},\ X_i \overset{iid}{\sim} B(p),\ 0 < p < 1.$$

$$\frac{1}{n} H_n(X) = H_1(X) = h(p) = p \log \frac{1}{p} + (1 - p) \log \frac{1}{1 - p}$$



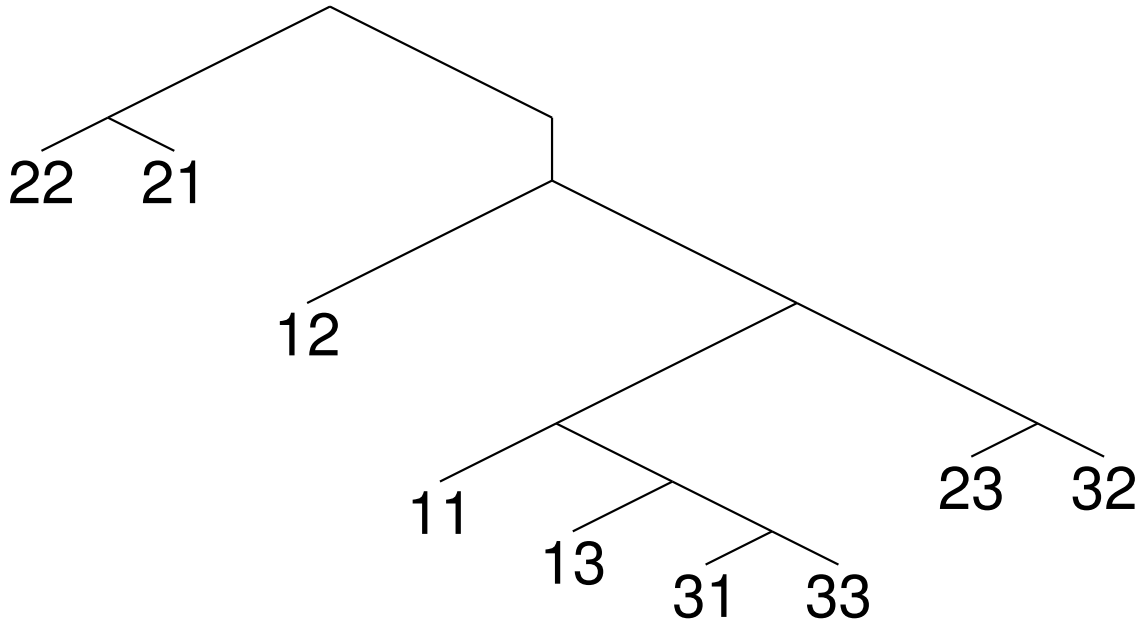
Aurélien Garivier
<garivier@enst.fr>

Example : a 2-block code

$A = \{1, 2, 3\}, n = 2, P^1 = 0.3\delta_1 + 0.6\delta_2 + 0.1\delta_3$ i.i.d.

$H(X) = -0.3 \times \log(0.3) - 0.6 \times \log(0.6) - 0.1 \times \log(0.1) \approx 1.295$ bits

X_1^2	$\phi_2(X_1^2)$
11	1100
12	10
13	11010
21	00
22	01
23	1110
31	110110
32	1111
33	110111



$$\frac{1}{2}\mathbb{E}\left[\left|\phi_2(X_1^2)\right|\right] \approx 1.335$$

Aurélien Garivier
<garivier@enst.fr>

Entropy Rate

- **Theorem**[Shannon-Breiman-McMillan] if P is stationary and ergodic, then

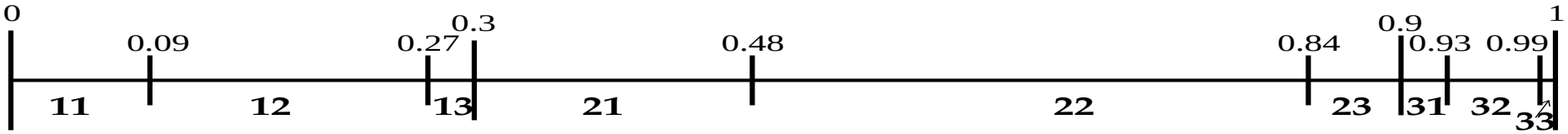
$$\frac{1}{n} H_n(X) \to H(X)$$

- $\Longrightarrow$ Interpretation : $H(X)$ = *minimal number of bits necessary to code each symbol of X*
- Kraft's inequality : for each code ϕ_n there exists a (sub-)probability distribution Q^n such that

$$Q^n(x) = 2^{-|\phi_n(x)|} \iff -\log Q^n(x) = |\phi_n(x)|$$

Arithmetic Coding : a constructive reciprocal

$$\phi_n(x_1^n) = \text{ the } \lceil -\log Q^n(x_1^n) \rceil + 2 \text{ first bits of } \frac{Q^n(X_1^n \leq x_1^n) + Q^n(X_1^n < x_1^n)}{2}$$



Ex : $A = \{1, 2, 3\}, n = 2,$
$Q^2 = [0.3, 0.6, 0.1]^{\otimes 2}$

- $\mathbf{I}(\mathbf{11}) = [\mathbf{0}, \mathbf{0.09}[$:

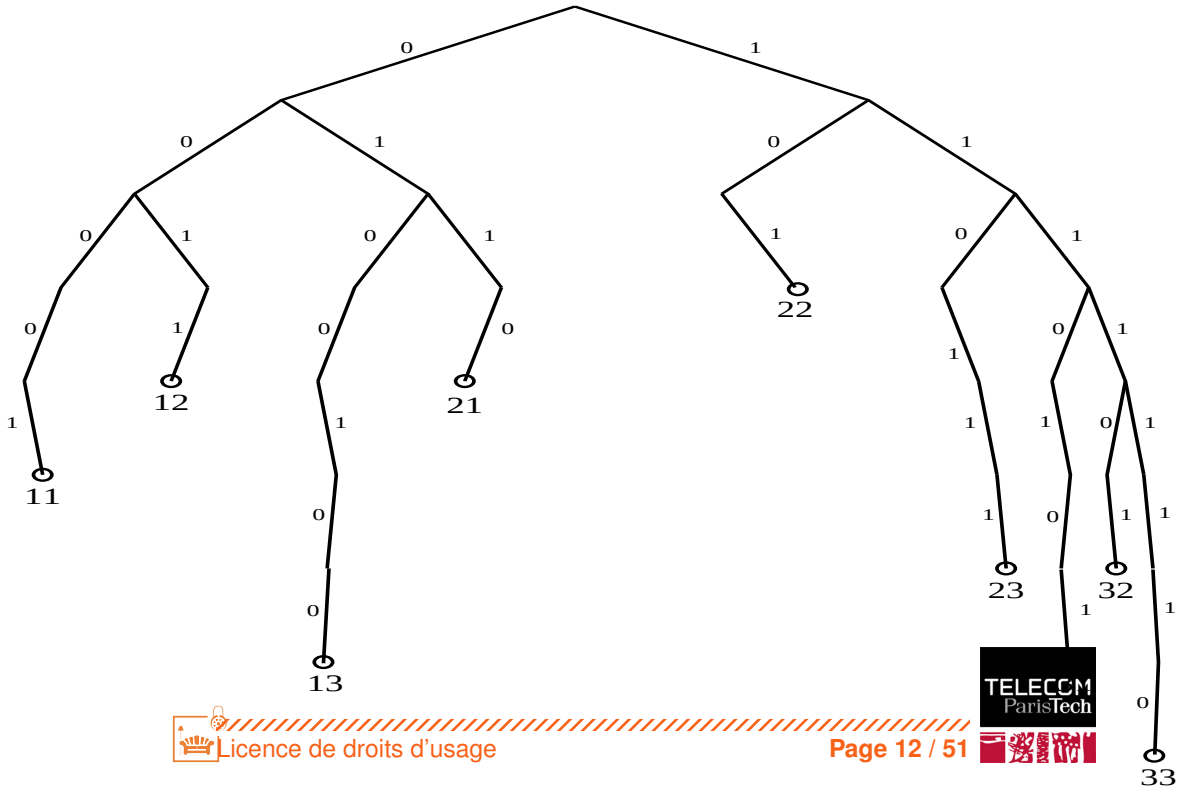
$m(00) = 0.045 = \overline{0.\mathbf{0000}101110\ldots}$

As $\left\lceil \log \frac{1}{0.09} \right\rceil + 1 = 5$, $\phi_2(00) = 00001$

- $\mathbf{I}(\mathbf{22}) = [\mathbf{0.48}, \mathbf{0.84}[$:

$m(22) = 0.66 = \overline{0.\mathbf{101}0100011\ldots}$

As $\left\lceil \log \frac{1}{0.36} \right\rceil + 1 = 3$, $\phi_2(22) = 101$



Coding distributions

$\Longrightarrow$ $-\log Q^n(x) =$ *code length* for x with *coding distribution* Q^n.

- Th Shannon : the best coding distribution is $Q^n = P^n$.
- Using another distribution causes an expected overhead called *redundancy*

$$\begin{aligned} \mathbb{E}_{\mathbb{P}}\left[|\phi_n(X_1^n)|\right] - H(X_1^n) &= \mathbb{E}_{\mathbb{P}}\left[-\log Q^n(X_1^n) + \log P^n(X_1^n)\right] \\ &= KL(P^n, Q^n) \end{aligned}$$

= Küllback-Leibler divergence of P^n to Q^n.

Universal coding

- What if the source statistics are unknown ?
- What if we need a versatile code ?

$\Longrightarrow$ Need a single coding distribution Q_n for a class of sources

$$\Lambda = \{\mathbb{P}_\theta, \theta \in \Theta\}$$

Ex : memoryless processes, Markov chains, HMM...

$\Longrightarrow$ unavoidable redundancy, in the worst case :

$$\sup_{\theta \in \Theta} \mathbb{E}_{P_\theta} \left[|\phi_n (X_1^n)| \right] - H(X_1^n) = \sup_{\theta \in \Theta} KL(P_\theta^n, Q^n)$$

$\Longrightarrow$ Q^n must be close to all the $\{P_\theta, \theta \in \Theta\}$ for KL divergence

Predictive Coding

- What if we want to code messages with different or large lengths ?
- Consistent coding family :

$$\forall x_1^n \in A^n, Q^n(x_1^n) = \sum_{a \in A} Q^{n+1}(xa)$$

- Define $Q(a|x_1^n) = \frac{Q(x_1^n a)}{\sum_{b \in A} Q(x_1^n b)}$, then

$$Q^n(x_1^n) = \prod_{j=1}^{n} Q\left(x_j | x_1^{j-1}\right)$$

- Predictive coding scheme : update $Q(\cdot|x_1^n)$
- Example : Laplace code $Q(a|x_1^n) = \frac{n(a,x_1^n)+1}{n+|A|}$, where $n(a, x_1^n) = \sum_{j=1}^{n} \mathbb{1}_{x_j=a}$.

Density estimation with Logarithmic Loss

- Feature space : context $x_1^{k-1} \in A^*$,
- Instant loss : number of bits used $\ell(Q, x_k) = -\log Q(x_k|x_1^{k-1})$
- Best regressor in average :
  $Q^*(\cdot|x_1^{k-1}) = \mathbb{P}_\theta\left(X_k = x_k|X_1^{k-1} = x_1^{k-1}\right)$
- Excess loss : $\ell(Q, x_k) - \ell(Q^*, x_k) = \log \frac{\mathbb{P}_\theta(X_k=x_k|X_1^{k-1}=x_1^{k-1})}{Q^k\left(x_k|x_1^{k-1}\right)}$

$$\mathbb{E}_{\mathbb{P}_\theta}\left[\ell(Q, x_k) - \ell(Q^*, x_k)\right] = KL\left(\mathbb{P}_\theta(X_k = \cdot|X_1^{k-1} = x_1^{k-1}), Q^k\left(\cdot|x_1^{k-1}\right)\right.$$

- Total excess loss : $L_n(Q) = \sum_{k=1}^n \ell(Q, x_k) - \ell(Q^*, x_k) = \log \frac{P_\theta^n(x_1^n)}{Q^n(x_1^n)}$
- Total expected excess loss : $\mathbb{E}_{\mathbb{P}_\theta} L_n(Q) = KL(P_\theta^n, Q^n)$

Outline

Aurélien Garivier
<garivier@enst.fr>

Bayesian predictive coding

- Let π be any prior on Θ, take

$$Q^n(x_1^n) = \int P_\theta^n(x_1^n)\pi(d\theta)$$

- Predictive coding distribution :

$$Q(a|x_1^n) = \int \mathbb{P}_\theta\left(x_{n+1} = a|X_1^n\right)\pi(d\theta|x_1^n)$$

- Conjugate priors $\Longrightarrow$ easy update rules for $Q(\cdot|x_1^n)$
- Idea : $\forall\theta \in \Theta, Q^n(x_1^n) \geq P_{\theta\pm\delta}^n(x_1^n)\pi(\theta \pm \delta)$

Example : i.i.d. classes

- A stationary memoryless source is parameterized by

$$\theta \in \mathcal{S}_{|A|} = \left\{ (\theta_1, \ldots, \theta_{|A|}) : \sum_{i=1}^{|A|} \theta_j = 1 \right\}$$

- Conjugate Dirichlet prior $\pi = \mathcal{D}(\alpha_1, \ldots, \alpha_{|A|})$.
- Posterior : $\Pi\left(d\theta | x_1^n\right) = \mathcal{D}(\alpha_1 + n_1, \ldots, \alpha_{|A|+n_{|A|}})$

$$\Longrightarrow \quad Q(a|x_1^n) = \frac{n_a + \alpha_a}{n + \sum_{j=1}^{|A|} \alpha_j}$$

- Laplace code = special case $\alpha_1 = \cdots = \alpha_{|A|} = 1$

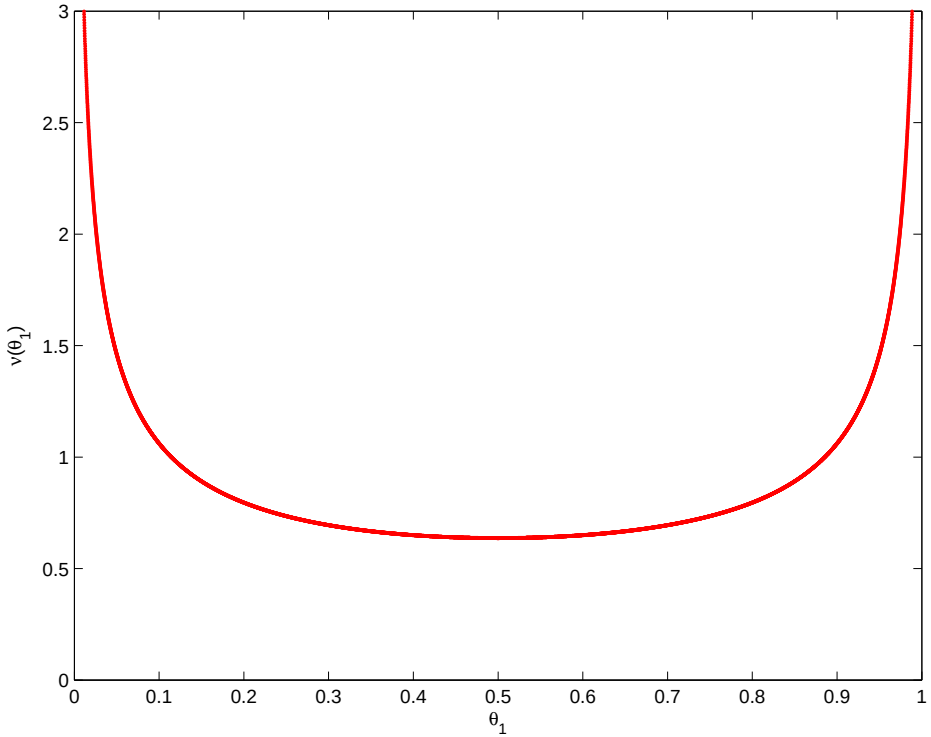
The Krichevski-Trofimov Mixture

Dirichlet prior with $\alpha_1 = \cdots = \alpha_{|A|} = 1/2$:

$$Q_{\mathrm{KT}}(x_1^n) = \int_{\theta \in \mathcal{S}_{|A|}} P_\theta(x_1^n) \, \frac{\Gamma\left(\frac{|A|}{2}\right)}{\sqrt{|A|}\Gamma\left(\frac{1}{2}\right)^{|A|}} \prod_{a \in A} \theta_a^{-1/2} d\theta_a.$$

- Jeffrey's prior = Non-informative
- Peaked on extreme parameter values
- Example |A|=2 :

$$\begin{aligned} & Q_{\mathrm{KT}}(001) \\ =& Q_{\mathrm{KT}}(0) Q_{\mathrm{KT}}(0|0) Q_{\mathrm{KT}}(1|00) \\ =& \frac{1/2}{1} \times \frac{1/2+1}{2} \times \frac{1/2}{3} \end{aligned}$$



Optimality

- **Theorem :** (Krichevsky-Trofimov '81) If Q^n_{KT} is the KT-mixture code and if $m = |A|$, then

$$\max_{\theta \in \Theta} KL(P^n_\theta, Q^n_{\mathrm{KT}}) = \frac{m-1}{2} \log \frac{n}{2} + \log \frac{\Gamma(1/2)^m}{\Gamma\left(\frac{m}{2}\right)} + o_m(1)$$

- **Theorem :** (Rissanen '84) If $\dim \Theta = k$, and if there exists a $\sqrt{n}$-consistent estimator of θ given X_1^n, then

$$\liminf_{n \to \infty} \min_{Q^n} \max_{\theta \in \Theta} KL(P^n_\theta, Q^n) \geq \frac{k}{2} \log n$$

- **Theorem :** Haussler'96
  there exists a Bayesian code Q^n reaching the minimax redundandy

Optimality : more precisely

- **Theorem :** (Krichevsky-Trofimov '81) If Q^n_{KT} is the KT-mixture code and if $m = |A|$, then

$$\max_{\theta \in \Theta} KL(P^n_\theta, Q^n_{\mathrm{KT}}) = \frac{m-1}{2} \log \frac{n}{2} + \log \frac{\Gamma(1/2)^m}{\Gamma\left(\frac{m}{2}\right)} + o_m(1)$$

- **Theorem :** (Xie-Barron '97)

$$\min_{Q^n} \max \theta \in \Theta KL(P^n_\theta, Q^n) = \frac{m-1}{2} \log \frac{n}{2e} + \log \frac{\Gamma(1/2)^m}{\Gamma\left(\frac{m}{2}\right)} + o_m(1)$$

- Actually, **pointwise** approximation :

$$-\log Q^n_{\mathrm{KT}}(x_1^n) \leq \inf_\theta -\log P^n_\theta(x_1^n) + \frac{|A|-1}{2} \log n + C$$

Sources with memory

- iid sources on finite alphabets are very-well treated...
- ...but they provide poor models for most applications !
- example : Shannon considered Markov models for natural languages
- Pb : a Markov Chain of order 3 on an alphabet of size 26 has $k = 439400$ degrees of freedom !

$$\min_{Q^n} \max_{\theta \in \Theta} KL(P_\theta^n, Q^n) \geq \frac{k}{2} \log n$$

$\Longrightarrow$ need of more *parcimonious* classes

Outline

Aurélien Garivier
<garivier@enst.fr>

Context Tree Sources

Formal definition :

- **Context Tree :** $T \subset A^*$ such that

$$\forall x_{-\infty}^{0}, \exists \text{ unique } s \in T : x_{-|s|}^{-1} = s$$

The unique suffix of $x_{-\infty}^{-1}$ in T is denoted by $T(x)$

- A string s is a **context of** a stationary ergodic process $\mathbb{P}$ if $\mathbb{P}\left(X_{-|s|}^{-1} = s\right) > 0$ and if for all $x_{-\infty}^{-|s|-1}$:

$$\mathbb{P}\left(X_0 = a \middle| X_{-\infty|}^{-1} = x_{-\infty}^{-|s|-1} s\right) = \mathbb{P}\left(X_0 = a \middle| X_{-|s|}^{-1} = s\right)$$

- T the **context tree of** $\mathbb{P}$ if T is a context tree, contains all contexts of $\mathbb{P}$ and if none of its elements can be replaced by a proper suffix without violating one of these properties

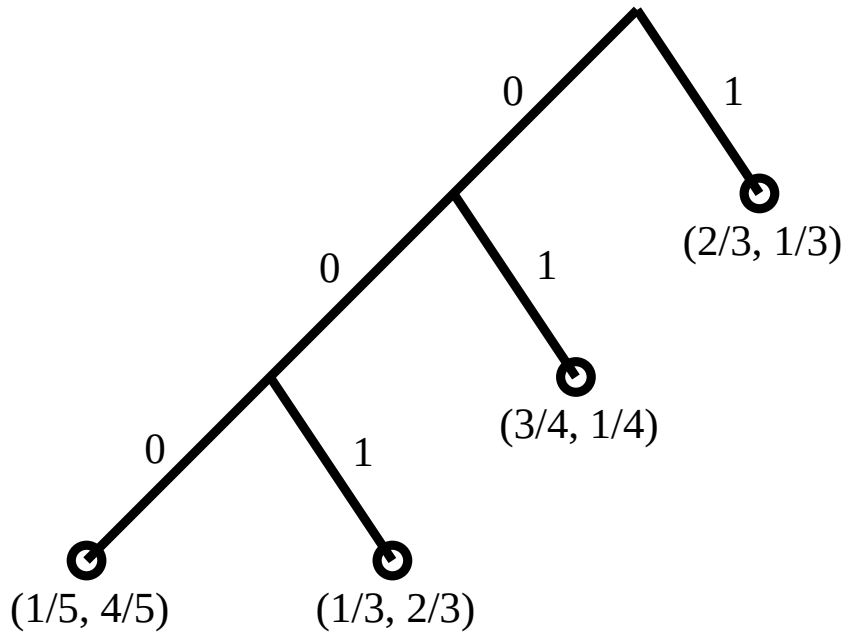
Context Tree Sources

Informal Definition A Context tree Source is a Markov Chain whose order is allowed to depend on the past data.

$$T = \{1, 10, 100, 000\}$$

$$P(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{aligned}
= \quad & P(X_1 = 0 | X_{-1}^0 = 10) && 3/4 \\
\times \quad & P(X_2 = 0 | X_{-1}^1 = 100) && 1/3 \\
\times \quad & P(X_3 = 1 | X_{-1}^2 = 1000) && 4/5 \\
\times \quad & P(X_4 = 1 | X_{-1}^3 = 10001) && 1/3 \\
\times \quad & P(X_5 = 0 | X_{-1}^4 = 100011) && 2/3
\end{aligned}$$



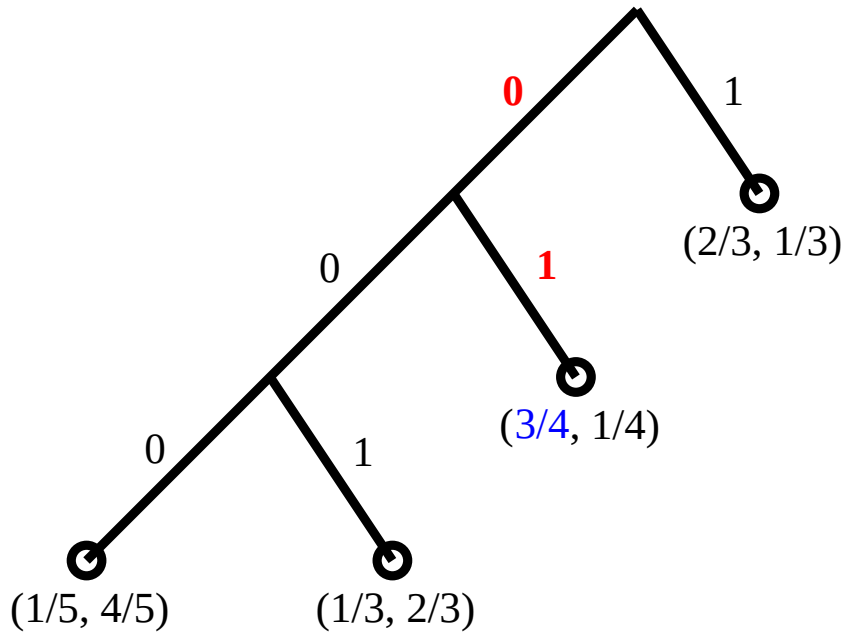
Context Tree Sources

Informal Definition A Context tree Source is a Markov Chain whose order is allowed to depend on the past data.

$$T = \{1, 10, 100, 000\}$$

$$P(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{aligned}
= \quad & P(X_1 = \mathbf{0} | X_{-1}^0 = 10) && 3/4 \\
\times \quad & P(X_2 = 0 | X_{-1}^1 = 100) && 1/3 \\
\times \quad & P(X_3 = 1 | X_{-1}^2 = 1000) && 4/5 \\
\times \quad & P(X_4 = 1 | X_{-1}^3 = 10001) && 1/3 \\
\times \quad & P(X_5 = 0 | X_{-1}^4 = 100011) && 2/3
\end{aligned}$$



Aurélien Garivier
<garivier@enst.fr>

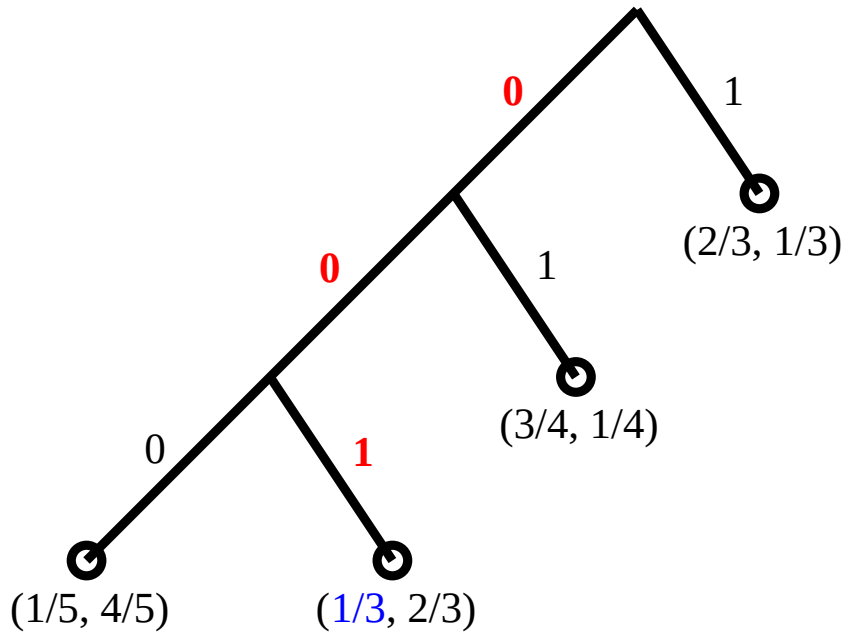
Context Tree Sources

Informal Definition A Context tree Source is a Markov Chain whose order is allowed to depend on the past data.

$$T = \{1, 10, 100, 000\}$$

$$P(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{aligned}
= \quad & P(X_1 = 0 | X_{-1}^0 = 10) && 3/4 \\
\times \quad & P(X_2 = \mathbf{0} | X_{-1}^1 = 100) && 1/3 \\
\times \quad & P(X_3 = 1 | X_{-1}^2 = 1000) && 4/5 \\
\times \quad & P(X_4 = 1 | X_{-1}^3 = 10001) && 1/3 \\
\times \quad & P(X_5 = 0 | X_{-1}^4 = 100011) && 2/3
\end{aligned}$$



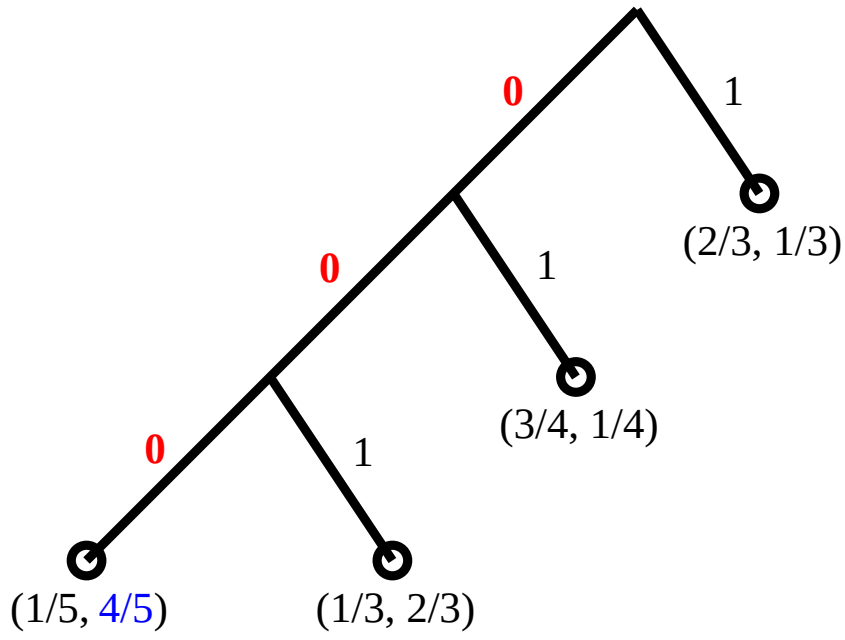
Context Tree Sources

Informal Definition A Context tree Source is a Markov Chain whose order is allowed to depend on the past data.

$$T = \{1, 10, 100, 000\}$$

$$P(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{aligned}
= \quad & P(X_1 = 0 | X_{-1}^0 = 10) && 3/4 \\
\times \quad & P(X_2 = 0 | X_{-1}^1 = 100) && 1/3 \\
\times \quad & P(X_3 = \mathbf{1} | X_{-1}^2 = 1000) && 4/5 \\
\times \quad & P(X_4 = 1 | X_{-1}^3 = 10001) && 1/3 \\
\times \quad & P(X_5 = 0 | X_{-1}^4 = 100011) && 2/3
\end{aligned}$$



Aurélien Garivier
<garivier@enst.fr>

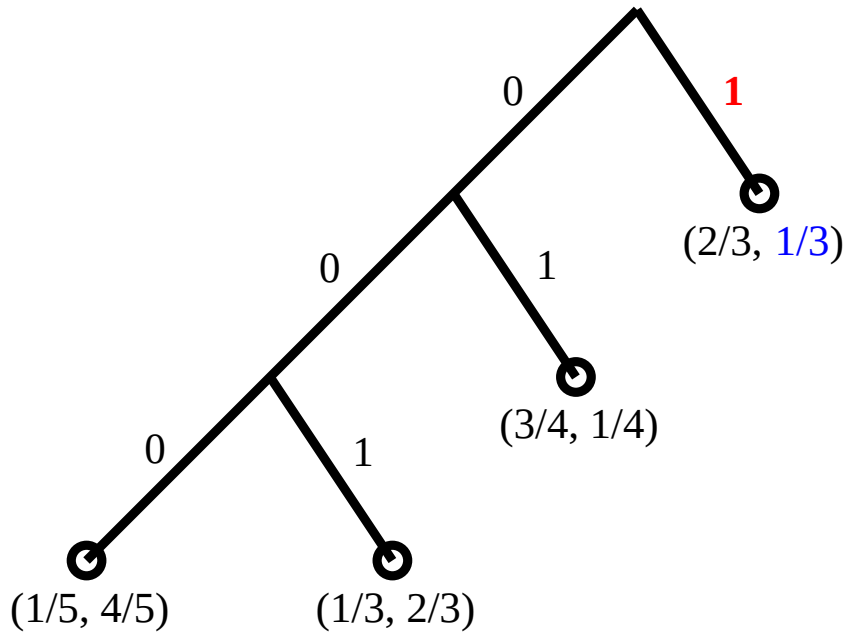
Context Tree Sources

Informal Definition A Context tree Source is a Markov Chain whose order is allowed to depend on the past data.

$$T = \{1, 10, 100, 000\}$$

$$P(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{aligned}
= \quad & P(X_1 = 0 | X_{-1}^0 = 10) && 3/4 \\
\times \quad & P(X_2 = 0 | X_{-1}^1 = 100) && 1/3 \\
\times \quad & P(X_3 = 1 | X_{-1}^2 = 1000) && 4/5 \\
\times \quad & P(X_4 = \mathbf{1} | X_{-1}^3 = 10001) && 1/3 \\
\times \quad & P(X_5 = 0 | X_{-1}^4 = 100011) && 2/3
\end{aligned}$$



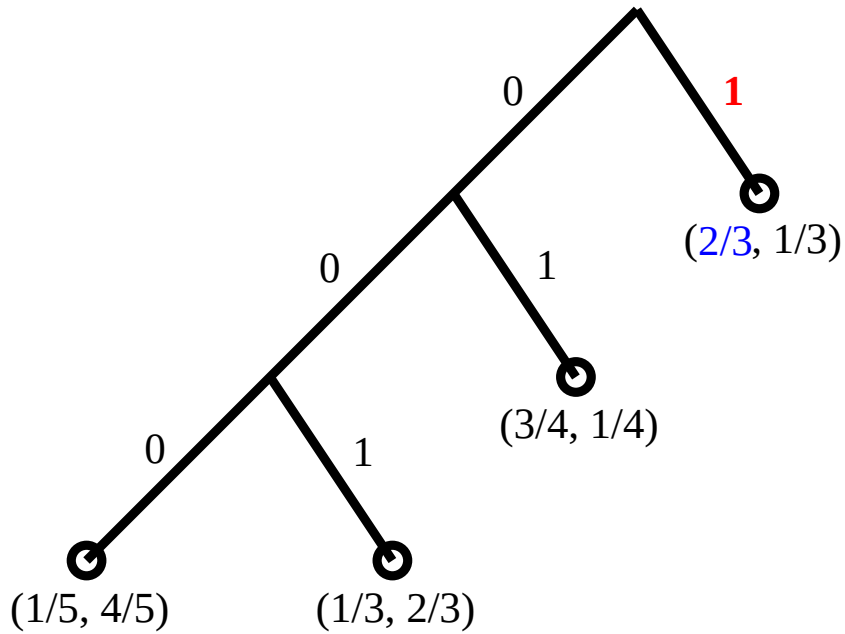
Context Tree Sources

Informal Definition A Context tree Source is a Markov Chain whose order is allowed to depend on the past data.

$$T = \{1, 10, 100, 000\}$$

$$P(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{aligned}
= \quad & P(X_1 = 0 | X_{-1}^0 = 10) && 3/4 \\
\times \quad & P(X_2 = 0 | X_{-1}^1 = 100) && 1/3 \\
\times \quad & P(X_3 = 1 | X_{-1}^2 = 1000) && 4/5 \\
\times \quad & P(X_4 = 1 | X_{-1}^3 = 10001) && 1/3 \\
\times \quad & P(X_5 = \mathbf{0} | X_{-1}^4 = 100011) && 2/3
\end{aligned}$$

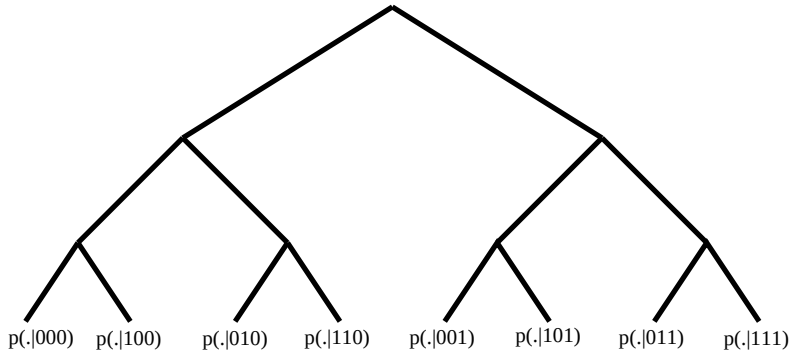


Context Trees versus Markov Chains

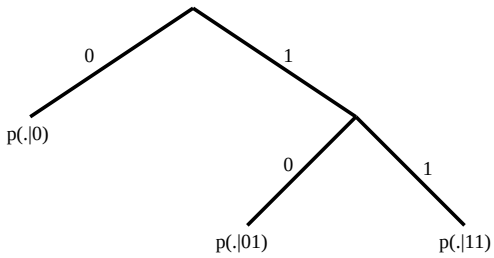
- Markov chain of order r = context tree source corresponding to a complete tree of depth r

  Markov chain of order 3

$$M = \begin{pmatrix} p(.|000) \\ p(.|100) \\ \vdots \\ p(.|111) \end{pmatrix} \Rightarrow$$



- Finite context tree source of depth d = Markov Chain of order d



$$\Rightarrow \quad M = \begin{pmatrix} p(.|0) \\ p(.|0) \\ p(.|01) \\ p(.|11) \end{pmatrix}$$

$\Longrightarrow$ much more flexibility : large number of models per dimensi

Aurélien Garivier
<garivier@enst.fr>

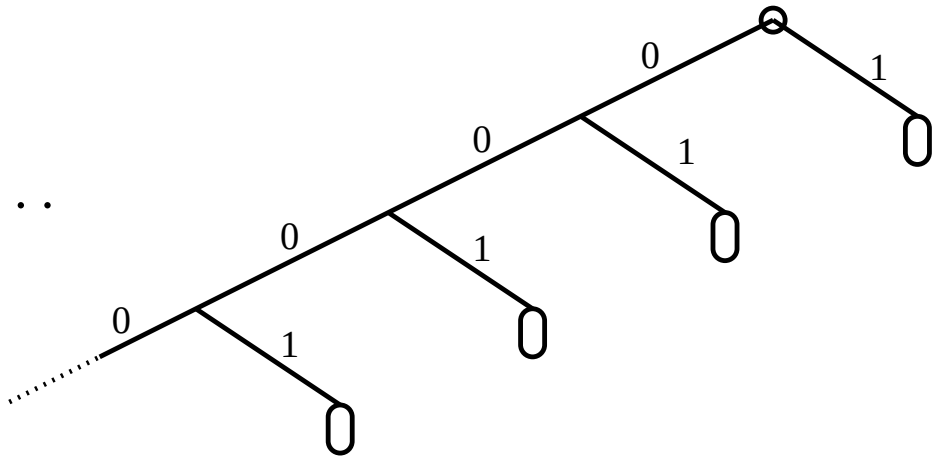
Infinite trees, unbounded memory

There exist (useful and simple) non-markovian context tree sources, see Galves, Ferrari, Comets,...

Ex : **Renewal Processes**

$$X_{-\infty}^{\infty} = \cdots 1 \underbrace{00\cdots1}_{N_i} \underbrace{00\cdots1}_{N_{i+1}} \cdots$$

$$N_i \overset{iid}{\sim} \mu \in \mathfrak{M}_1(\mathbb{N}_+)$$



$$\mathbb{P}\left(X_0 = 1 | X_{-\infty}^{-1} = \ldots 10^{j-1}\right) = \frac{\mu(j)}{\mu\left([j,\infty[\right)}$$

Outline

Aurélien Garivier
<garivier@enst.fr>

Bayes Codes in a Context Tree Model

- For a given context tree T, let

$$\Lambda_T = \{\mathbb{P} \text{ stationary ergodic on } A : \mathcal{T}(\mathbb{P}) = T\}$$

- $\mathbb{P} \in \Lambda_T$ is fully charactarized by $\theta = (\theta_s)_{s \in T} \in \Theta_T = \mathcal{S}_{|A|}^T \sim \mathbb{R}^{|T|(|A|-1)}$:

$$\mathbb{P}\left(X_0 = \cdot | X_{-\infty}^{-1} = x_{-\infty}^{-1}\right) = \theta_{T(x_{-\infty}^{-1})}(\cdot)$$

- Bayesian coding : prior π_T on Θ_T,

$$Q_{\pi_T}(x) = \int \mathbb{P}_\theta(x_1^n) d\pi_T(\theta)$$

Krichevsky-Trofimov Mixtures for Context Trees

- Prior : the θ_s are independent, $\mathcal{D}\left(\frac{1}{2},\ldots,\frac{1}{2}\right)$- distributed :

$$\pi_T(d\theta) = \bigotimes_{s\in T} \pi_{KT}(d\theta_s)$$

- For $s \in T$ let $T(x_1^n, s)$ = substring of x_1^n that occurs in context s
  Example : for $T = \{00, 10, 1\}$ and $x_0^8 = 1\ 00100001011$,

  $T(x_1^n, 00) = 0101 \quad T(x_1^n, 10) = 001 \quad T(x_1^n, 1) = 0001$
- The KT-mixture on model T is :

$$Q_{\mathrm{T}}^n(x_1^n) = \prod_{s\in T} Q_{\mathrm{KT}}^{|T(x_1^n,s)|}\left(T(x_1^n, s)\right)$$

Optimality

- **Theorem :** (Willems, Shtar'kov, Tjalkens '93) there is a constant C such that :

$$-\log Q_{\mathrm{T}}^{n}(x_1^n) \leq \inf_{\theta \in \mathcal{S}_{|A|}^{T}} -\log \mathbb{P}_\theta(x_1^n | x_{-\infty}^{-1}) + |T| \frac{|A|-1}{2} \log\left(\frac{n}{|T|}\right) + C|T|$$

- first-order optimal
- What if we don't know which model T to use ?

$$\Longrightarrow \quad \text{model selection}$$

Minimum Description Length Principle

- Occam's razor (Jorma Rissanen '78) :

  Choose the model that gives the shortest description of data

- Information theory :

  objective codelength = codelength of a minimax coder.

- Estimator associated with minimax mixtures $(\pi_T)_T$:

$$\hat{T}_{BIC} = \arg\min_T -\log \int_{\theta\in\Theta_T} P_\theta^n(x_1^n)\,\pi_T(d\theta)\,.$$

- Looking directly at the bounds :

$$\hat{T}_{KT} = \arg\min_T \inf_{\theta\in\Theta_T} -\log P_\theta^n(x_1^n) + \frac{|T|(|A|-1)}{2}\log n.$$

= penalized maximum likelihood estimator with BIC penalty.

Consistency Results

Theorems (Csiszár& Shields '96, Csiszár& Talata '04, G. '05, Galves & Leonardi '07, Leonardi '09) :

- the BIC estimator is strongly consistent
- the Krichevski-Trofimov estimator is not consistent

Many Rates of convergence, minimal penalties, ...
The Krichevski-Trofimov estimator fails to identify Bernoulli-1/2 iid sources !
One can derive from this bounds on $\sup_T \sup_{\theta \in \Theta_T} KL\left(P_\theta^n, Q_{\hat{T}}^n\right)$

Outline

Aurélien Garivier
<garivier@enst.fr>

Double mixture

- ν = measure on the set $\mathcal{T}$ of all context trees defined by :

$$\nu(T) = 2^{-2|T|+1}$$

  is a **probability distribution**.

- **Context Tree Weighting** coding distribution :

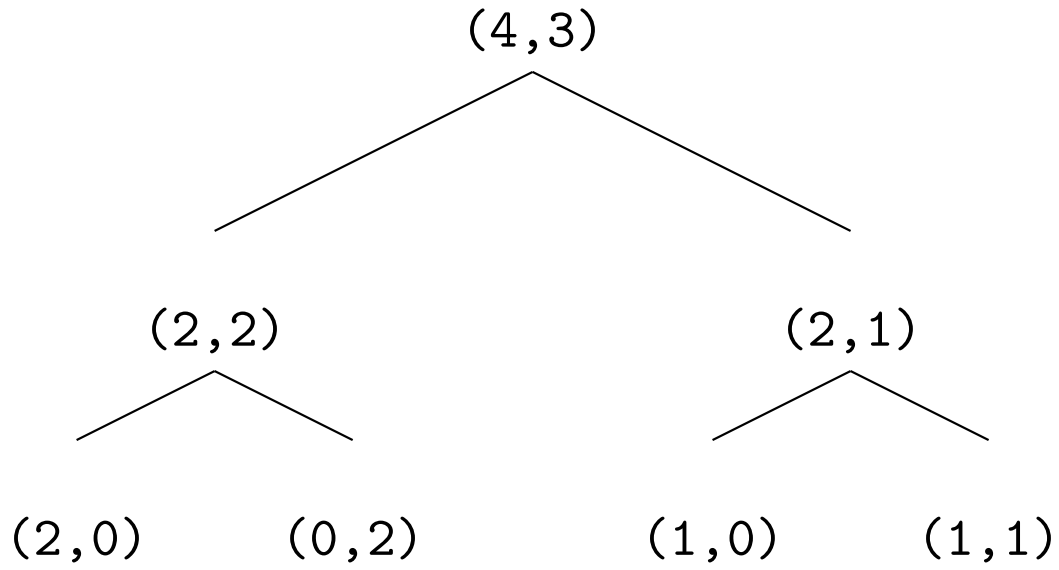
$$Q^n_{\mathrm{CTW}}(x_1^n) = \sum_T \nu(T) Q^n_{\mathrm{T}}(x_1^n)$$

- **Theorem** (Shtarkov and al '93)) : oracle inequality :

$$-\log Q^n_{\mathrm{CTW}}(x_1^n | x_{-\infty}^0) \leq \inf_{T \in \mathcal{T}} \inf_{\theta \in \Theta_T} p_\theta(x_1^n | x_{-\infty}^0) + \frac{|A|-1}{2} |T| \log^+ \frac{n}{T} + |T|(2 + \log m) + m - 2.$$

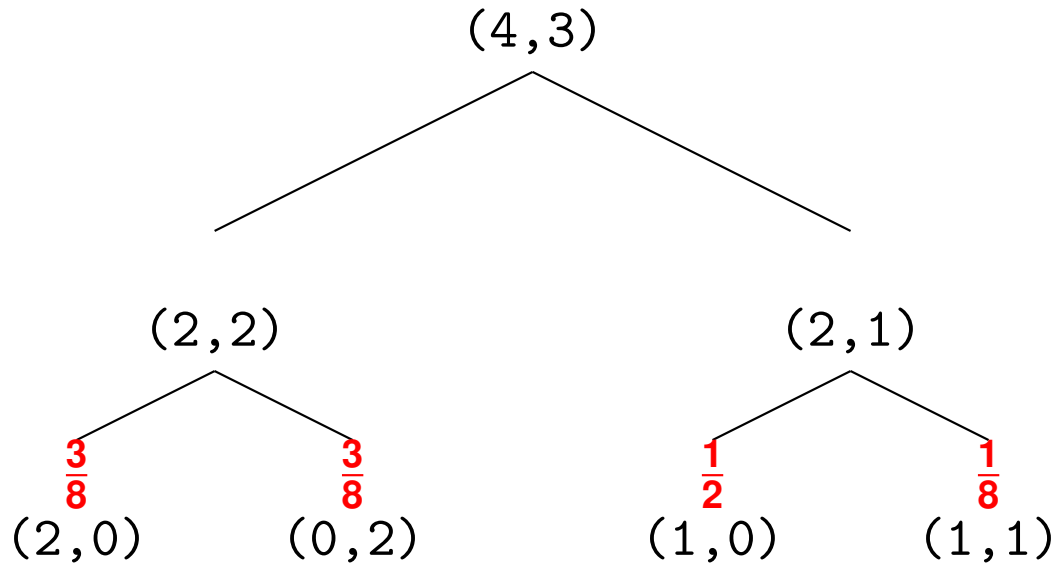
CTW : Efficient Algorithm

Idea : in each node, compute the arithmetical mean of a *self-prob* and a *sub-prob*.



CTW : Efficient Algorithm

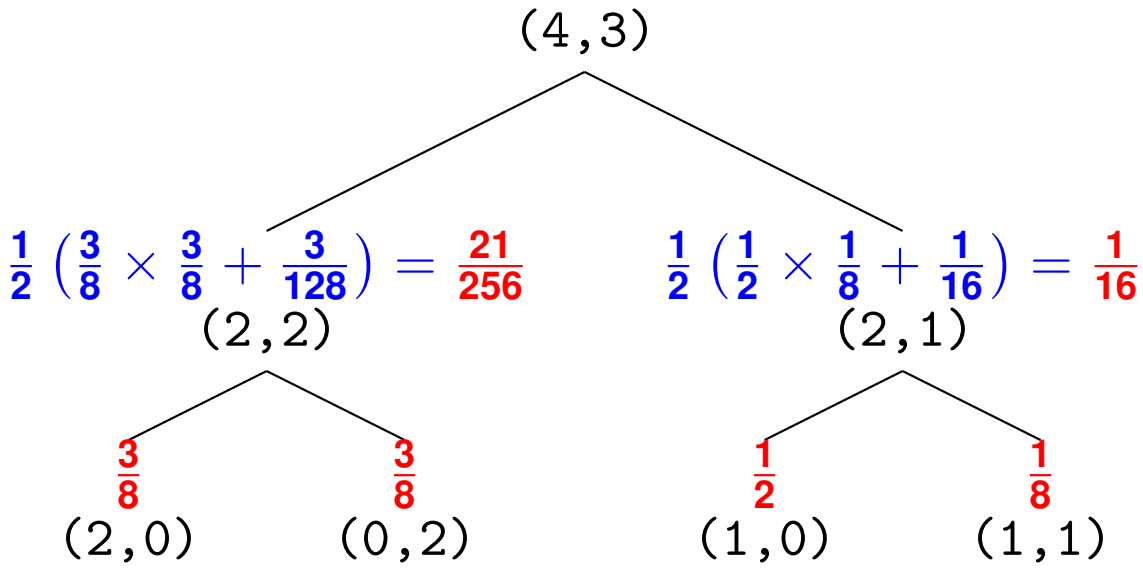
Idea : in each node, compute the arithmetical mean of a *self-prob* and a *sub-prob*.



Aurélien Garivier
<garivier@enst.fr>

CTW : Efficient Algorithm

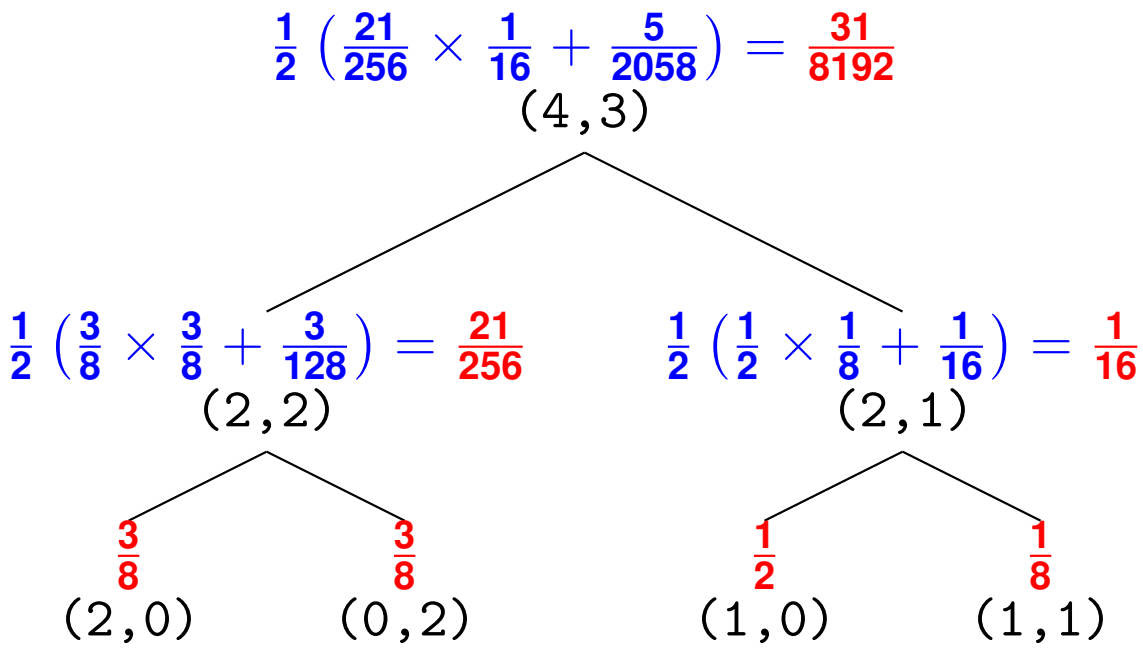
Idea : in each node, compute the arithmetical mean of a *self-prob* and a *sub-prob*.



Aurélien Garivier
<garivier@enst.fr>

CTW : Efficient Algorithm

Idea : in each node, compute the arithmetical mean of a *self-prob* and a *sub-prob*.



Aurélien Garivier
<garivier@enst.fr>

Outline

Aurélien Garivier
<garivier@enst.fr>

Renewal Processes

- Let $\mathcal{R}$ be the (non-parametric) class of renewal processes on the binary alphabet $\{0,1\}$:

$$X = \cdots 1 \underbrace{00\cdots1}_{N_i} \underbrace{00\cdots1}_{N_{i+1}} \cdots, \qquad N_i \overset{iid}{\sim} \mu \in \mathfrak{M}_1(\mathbb{N}_+)$$

- **Theorem :** (Csiszár and Shields '96) There exist two constants c and C such that

$$c\sqrt{n} \leq \inf_{Q^n} \sup_{\mathbb{P}\in\mathcal{R}} KL(P^n, Q^n) = C\sqrt{n}$$

$\Longrightarrow$ first example of an *intermediate complexity* class.

- Problem : non-constructive
- Does a *general purpose* coder behave well on Renewal Processes ?

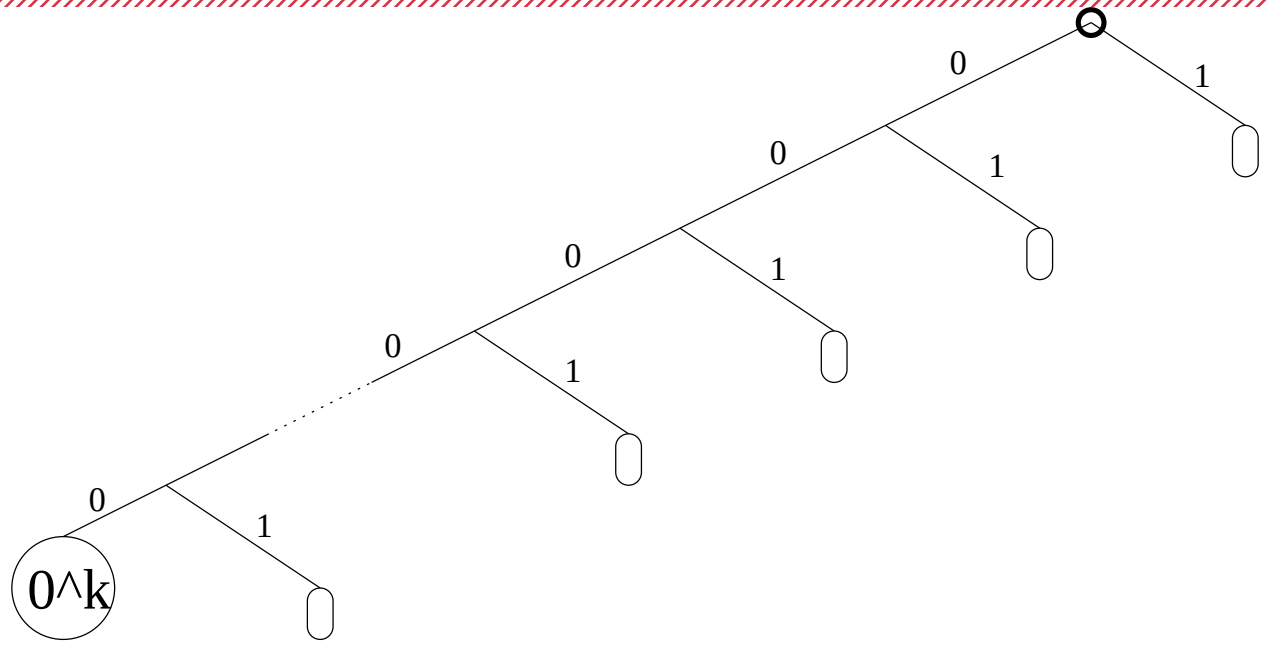
Redundancy of the CTW method on Renewal Processes

Theorem : (G. '04)
There exist constants c and C such that for all $\mathbb{P} \in \mathcal{R}$:

$$c\sqrt{n}\log n \leq KL\left(\mathbb{P}^n, Q^n_{\mathrm{CTW}}\right) \leq C\sqrt{n}\log n$$

- CTW is efficient on long-memory processes
- *Adaptivity :* if the renewal distribution is bounded, CTW achieves regret $O(\log n)$ (contrary to ad-hoc coders).
- Requires deep contexts in the double mixture ($\Longrightarrow$ the tree should not be cut off at depth $\log n$).
- Kind of non-parametric estimation : need for a balance between approximation and estimation.

Why it works



$\mathbb{P}^n$ is approximated by Q_{T}^n with T as above. Two losses :

- estimating the transition parameters : $k \log(n)/2$
- approximation due to boundedness : $n \log(k)/k$

$$\Longrightarrow \quad \text{for } k = c\sqrt{n}, \text{ both are } O\left(\sqrt{n}\log(n)\right)$$

Extension : Markovian Renewal Processes

- A *Markovian Renewal Processes* is such that successive distances between occurences of symbol 1 form a Markov Chain on $\mathbb{N}_+$:

$$X = \cdots 1 \underbrace{00\cdots1}_{N_i} \underbrace{00\cdots1}_{N_{i+1}} \cdots, \qquad (N_i)_i \text{ Markov chain}$$

- **Theorem :** (Csiszár and Shields '96) There exist two constants c and C such that

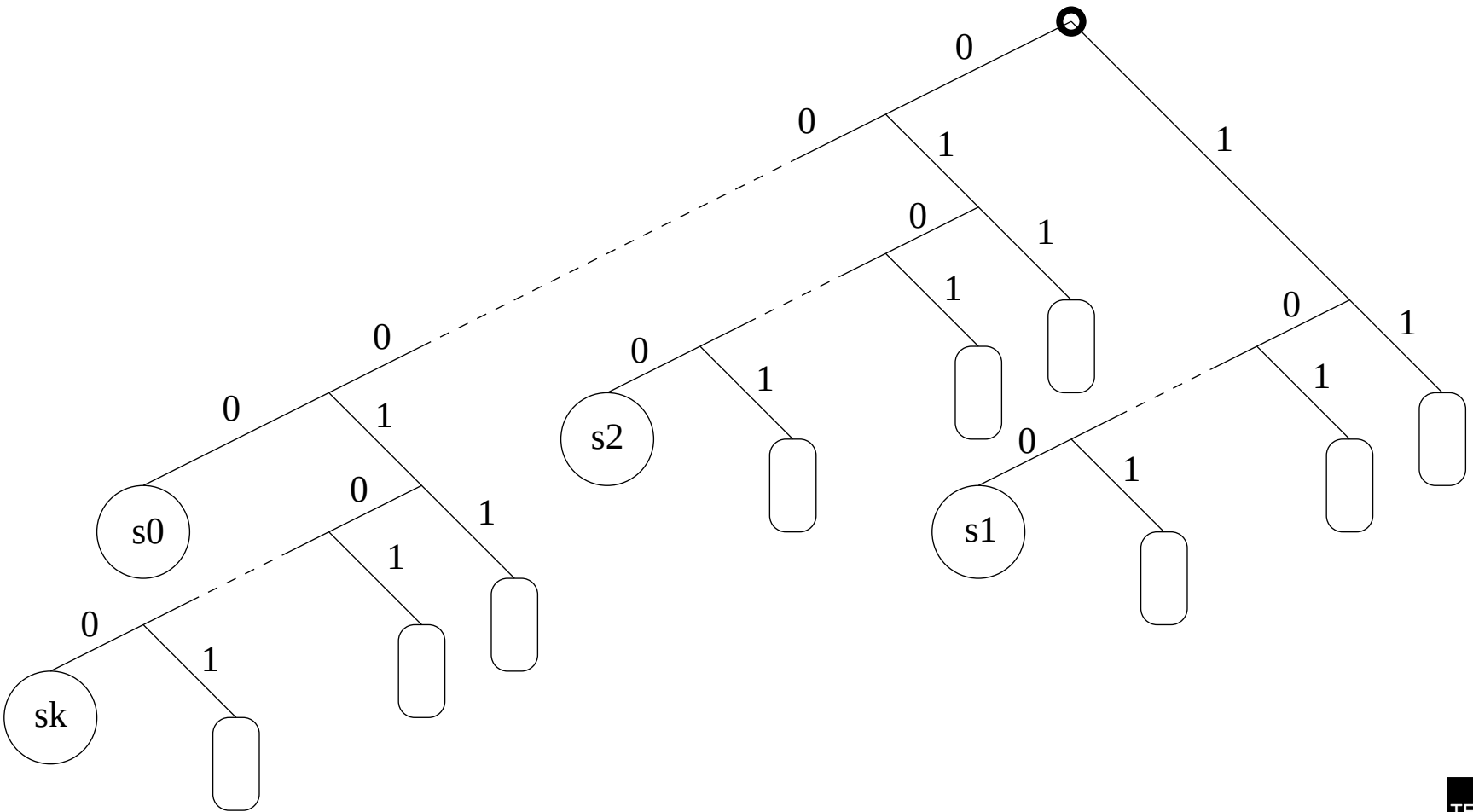
$$cn^{2/3} \leq \inf_{Q^n} \sup_{\mathbb{P}\in\mathcal{MR}} KL(\mathbb{P}^n, Q^n) \leq Cn^{2/3}$$

- **Theorem :** (G. '04) CTW is almost adaptive on the class $\mathcal{MR}$ of Markovian Renewal processes : $\exists C \in \mathbb{R}, \forall \mathbb{P} \in \mathcal{MR}$,

$$KL\left(P^n, Q^n_{\mathrm{CTW}}\right) \leq Cn^{2/3} \log n$$

Markovian Renewal Processes are Context Tree Sources

And so are r-order Markov Renewal Processes



Outline

Aurélien Garivier
<garivier@enst.fr>

Perspectives : oracle inequalities on more general classes ?

- CTW is probably more useful than expected : it obtains good results on infinite memory processes
- We want to obtain oracle inequalities for different non-markovian classes
- But the setting is more difficult than usually
- And approximation theory is difficult !
  Ex : let $\mathbb{P}_H$ be a HMM with 3 hidden states, what is

$$\inf_{|T|\leq k} \inf_{\theta \in \mathcal{S}^T_{|A|}} KL(P^n_H, P^n_\theta) \qquad ???$$

the end. . .

Thank you for your attention !